ISIP 2000 CONVERSATIONAL SPEECH EVALUATION SYSTEM

R. Sundaram, A. Ganapathiraju, J. Hamaker and J. Picone

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University, Mississippi State, MS 39762
{sundaram, ganapath, hamaker, picone}@isip.msstate.edu

ABSTRACT

In this paper, we describe the ISIP Automatic Speech Recognition system (ISIP-ASR) used for the Hub-5 2000 English evaluations. The system is a public domain cross-word context-dependent HMM based system and has all the functionality normally expected in an LVCSR system, including Baum-Welch training for continuous density HMMs, phonetic decision tree-based state-tying, word graph generation and rescoring. The acoustic models were trained on 60 hours of Switchboard and 20 hours of CallHome data. The system had a word error rate of 43.4% on Switchboard, 54.8% on CallHome, and an overall error rate of 49.1%. This paper describes the evaluation system in detail and discusses our post-evaluation experiments and improvements.

1. SYSTEM OVERVIEW

The ISIP-ASR system is a public domain cross-word context-dependent HMM-based system that is freely available for both commercial and academic use with no licensing or copyright restrictions [1]. It consists of three primary components: the acoustic front-end, HMM parameter estimation module and a hierarchical single-pass Viterbi decoder. Acoustic training has been enhanced to incorporate both Viterbi and Baum-Welch algorithms. The decoder can perform N-gram decoding and process word graphs.

1.1. Acoustic Front-End

The system uses a common front-end that transforms the input speech signal into mel-spaced cepstral coefficients appended with their first and second derivatives [2]. Standard features of this front-end are pre-emphasis filtering, windowing, debiasing, and energy normalization. To improve robustness to channel variations and noise, our evaluation system incorporated side-based cepstral mean subtraction [3]. Cepstral mean subtraction is computed as follows:

$$y_k(t) = x_k(t) - \bar{x}_k(t) \quad (1)$$

k = 1, 2,..., N where k is the cepstral index. $\bar{x}_k(t)$ is an estimate of the mean computed from all analysis frames belonging to the same conversation side as x_k.

For the evaluation, we used the front-end to generate 12 FFT-derived cepstral coefficients and log-energy. These features were computed using a 10 ms analysis frame and a 25 ms Hamming window. First and second derivative coefficients of the base features are appended to produce a thirty-nine dimensional feature vector. The 12 base cepstral features are then debiased using side-based cepstral mean subtraction.

1.2. Parameter Estimation

The training module consists of an Expectation Maximization (EM) based acoustic optimizer which uses the Baum-Welch algorithm for robust parameter estimation. This parameter estimation component supports continuous-density Gaussian mixture models with diagonal covariances. It also supports context-dependent models with state and model tying.

A problem often associated with training context-dependent models is the lack of training data to cover all the models in the system. To avoid this problem maximum likelihood phonetic decision tree-based state-tying is employed in the system [4]. The decision tree uses phonetic rules that are based on left and right contexts and a tree is grown for each state of each context-independent phone in the system. The evaluation system uses a context of one phone on either side of the center phone. The states of models with similar phonetic contexts are allowed to share data by tying them together. This leads to better parameter estimates as all of the model clusters are

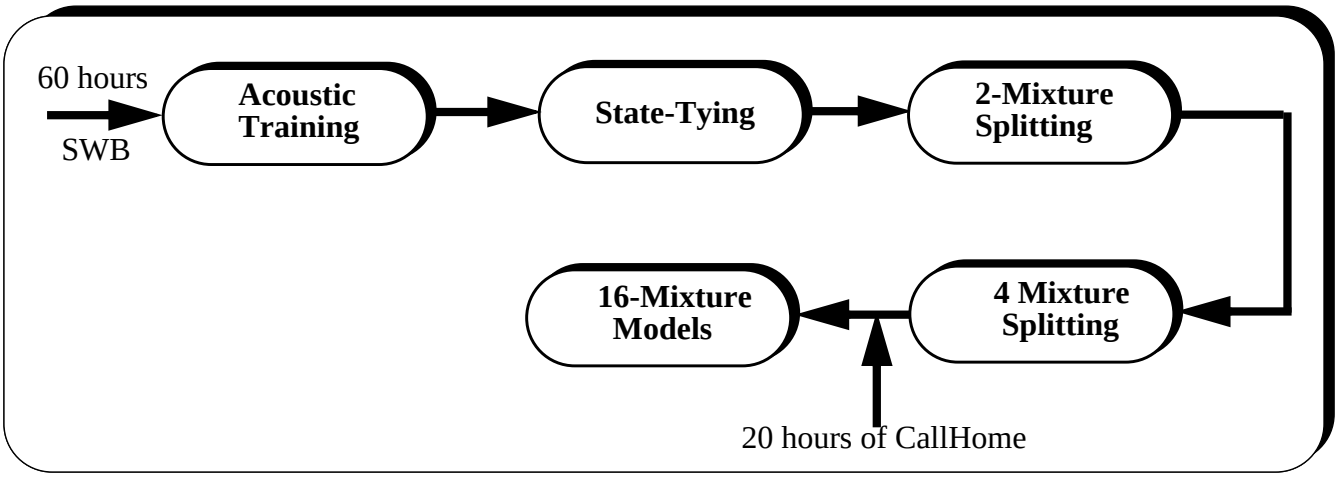

Figure 1: Data flow for acoustic model training. Note that the CallHome data is only incorporated into the training process after 4-mixture training.

seen in the training set a sufficient number of times. This also allows the system to generate models for unseen contexts. Table 1 shows the reduction in unique HMM states due to state-tying.

A synopsis of the acoustic model estimation in our 2000 evaluation system is shown in Figure 1. The system was trained on 60 hours of Switchboard-I data from 2998 conversation sides, and 20 hours of CallHome data from 240 conversation sides. Context-independent (CI) phone models were first trained using only the Switchboard data. These CI models were iteratively trained from one mixture component to 32 mixture components, and were then used to generate phone-level alignments. These alignments were used throughout the remainder of our training process. Context-dependent phone models were seeded with single-mixture monophones, reestimated using a four pass procedure, and then state-tied to cluster those states and models that were statistically similar. Mixture splitting was done using an iterative splitting and training scheme. After the four-mixture models were trained, CallHome data was added to the training set and the training continued to finally generate 16 mixture models. Word-internal and cross-word context-dependent phone models were built in this process.

System	Number of States	
	Before State-tying	After State-tying
Word-Internal	9580	4194
Cross-Word	67684	10619

Table 1: Number of states in the system before and after state-tying.

1.3. Language Model and Lexicon

We used both bigram and trigram backoff language models in the evaluation system. The language models were provided by SRI and were trained by interpolating language models generated using Switchboard, CallHome and Broadcast News (BN) data. The bigram version was used to generate word graphs while the trigram LM was used for rescoring. The trigram and bigram LM's were pruned using SRI's entropy-based method [5] to eliminate negligible bigram and trigram parameters. The final trigram language model contained 138k trigrams, 320k bigrams and 33k unigrams. The final bigram LM was obtained by all trigrams from the trigram LM.

The lexicon used by the system had a vocabulary of 22,000 words derived from the WS'97 test lexicon. This lexicon was then expanded to include words present in the SRI language model but not present in our original lexicon. The final lexicon had a vocabulary of 33,200 entries.

1.4. Recognition

The ISIP-ASR decoder is based on a hierarchical implementation of the standard time-synchronous

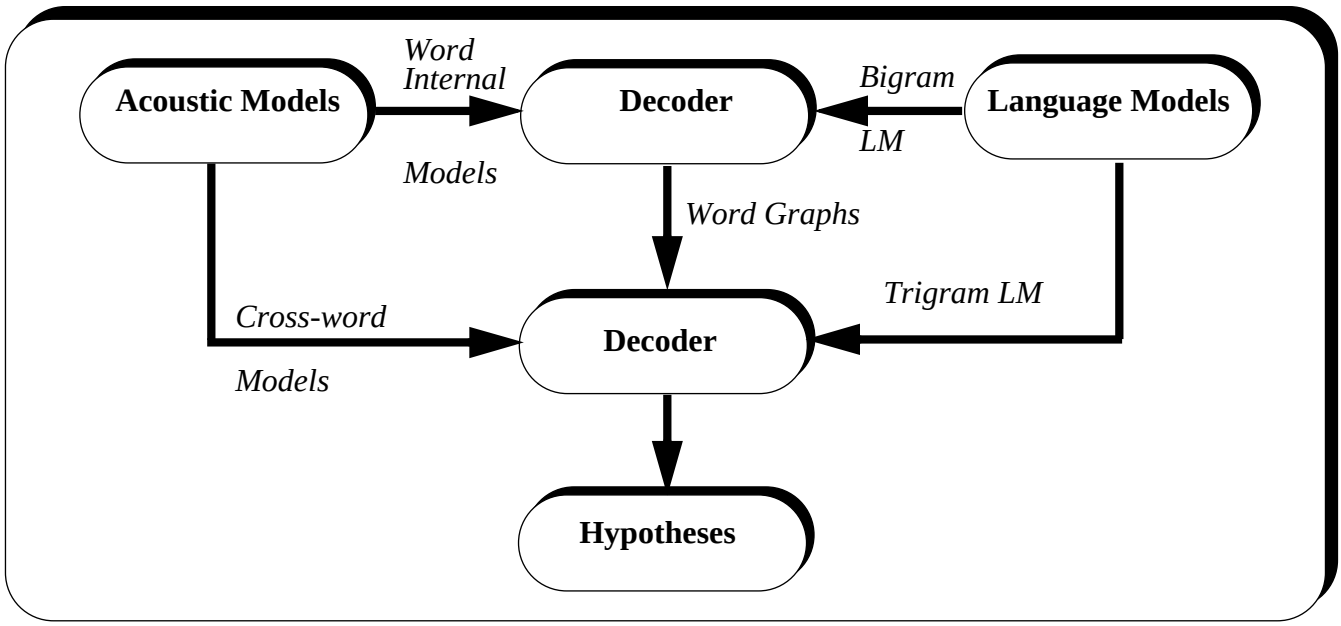

Figure 2: A two-stage evaluation system. It consists of a pre-processing stage where word-internal acoustic models and a bigram language model are used to generate word graphs. A second stage processes the word graphs while using cross-word acoustic models and a trigram language model

Viterbi search paradigm [6]. The decoder supports various modes such as N-gram decoding, word graph generation, word graph rescoring and supervised alignment. The decoder can handle both word-internal and cross-word context-dependent models, and uses a lexical tree-based organization to conserve memory during context expansion. Pruning techniques are employed at all levels in the search space to improve computational efficiency without significantly increasing error rate.

For the evaluation, recognition was performed in two stages using the decoder as shown in Figure 2. In the first pass, we used 16-mixture word-internal context-dependent phone models and a bigram language model to generate word graphs. This stage was followed by word graph rescoring using 16-mixture cross-word context-dependent phone models and a trigram language model. The output of this pass was the final evaluation hypothesis.

Processing was performed on 600 MHz Pentium III machines running Solaris 7. These machines had 1 Gigabyte of main memory and 2 Gigabytes of swap space. The time and memory requirements for the two decoding stages are tabulated in Table 2.

2. RESULTS AND ANALYSIS

The performance of the system on the evaluation 2000 test set is shown in Table 3. We also conducted several experiments after the evaluations to analyze system performance, and to improve deficiencies. The details of this work are described below.

2.1. Error Analysis

From Table 3 it can be noted that the overall system performance is better on Switchboard than on CallHome. This is consistent with other published results. However, the gap between our Switchboard results and CallHome results is larger than normal and may be due to the late introduction of the CallHome data into the training procedure.

An analysis of the word graphs produced in the first

	Time (hrs)	Memory (MB)
Pass 1	1083	500
Pass 2	23	300

Table 2: Time and memory requirements for ISIP-ASR system on the entire evaluation dataset.

	Total	Male	Female
Call Home	54.8%	55.0%	54.8%
Switchboard	43.4%	41.5%	45.3%
Overall	49.1%	51.2%	45.4%

Table 3: Performance of ISIP-ASR system.

stage of our decoding process shows that poor quality of the word graphs (WER of 19.8%) is a major contributor to our high evaluation error rate. Note that this is only marginally better than the best error rate reported in the evaluation. We believe that this is in part caused by the heavy pruning of the bigram language model used to build the word graphs and also by the heavy pruning we employed during word graph generation. This hypothesis is currently being tested.

2.2. Post-evaluation Experiments

Table 3 also points out a difference in performance between male and female speakers. Performance of male speakers is better than that of female speakers, suggesting that our models were more tuned to male speakers. This is most likely a by-product of the fact that even though the number of males and females was approximately equal in the training data, the male speakers accounted for a larger percentage of the acoustic data. Our initial experiment with gender-dependent models has produced a 0.6% improvement in WER.

After the formal evaluations we performed a series of experiments to incorporate features into our system that are common in the other evaluation systems. We also found a serious algorithmic error in the way we handle N-gram language models during rescoring of word graphs with a trigram LM. Surprisingly, fixing this error gave only a 0.5% improvement in WER on the Switchboard portion of the evaluation set. We are currently investigating whether this problem was also an issue during the word graph generation stage of the decoding process.

3. CONCLUSION

The ISIP-ASR system has been used for the Hub-5 evaluations for the first time this year. The system configuration included multiple Gaussian mixture cross-word acoustic models with parameter sharing. The acoustic models were trained on 60 hours of Switchboard and 20 hours of CallHome. Recognition was done using a two pass strategy — first pass of word graph generation with word-internal models and a bigram LM followed by a second pass of word graph rescoring using cross-word acoustic models and a trigram LM. This system had a word error rate of 43.4% and 54.8% on the Switchboard and CallHome components of the evaluation dataset. A priority for our future work will be the introduction of our new search engine that accommodates large language models, and incorporation of a generalized acoustic modeling component that handles arbitrarily-sized context-dependent phone models.

4. ACKNOWLEDGEMENTS

We wish to thank Dr. Andreas Stolcke and SRI for their assistance in providing the language models for our evaluation system.

REFERENCES

[1] N. Deshmukh, et al, “A Public Domain Speech-to-Text System,” *Proc. Eurospeech*, vol. 5, Budapest, Hungary, Sept. 1999.

[2] J. Picone, “Signal Modeling Techniques in Speech Recognition,” *IEEE Proceedings*, vol. 81, no. 9, pp. 1215-1247, Sept. 1993.

[3] Heab-Umbach, et al, “Acoustic Modeling in the Philips Hub-4 Continuous-Speech Recognition System,” *Proc. DARPA BNTUW Workshop*, Lansdowne, VA, USA, Feb. 1998.

[4] S.J. Young, et al, “Tree-Based Tying For High Accuracy Acoustic Modeling,” *Proc. ARPA Workshop on Human Lang. Tech.*, pp. 286-291, Plainsboro, NJ, USA, Sept. 1994

[5] A. Stolke, “Entropy-Based Pruning of Backoff Language Models,” *Proc. DARPA BNTUW*, Lansdowne, VA, USA, pp. 270-274, Feb. 1998.

[6] N. Deshmukh, et al, "Hierarchical Search for Large Vocabulary Conversational Speech Recognition,” *IEEE Signal Proc. Mag.*, vol. 16, no. 5, pp. 84-107, Sept. 1999.